



Original Research Paper

Federated Deep Learning for Privacy-Preserving Multi-Site Liver Fibrosis and Steatosis Prediction: A Comprehensive Study on the BEHSOF Tabular Clinical Dataset

M. Kaliappan¹, S. Vimal², Mariappan E³, Preethy Rebecca⁴, S. Senthil⁵, V. Manimaran⁶, Rattan Singh⁷, Gaurav Dhiman⁸

¹Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: kaliappan@ritrjpm.ac.in).

²Department of Computer Science and Engineering, KIT- Kalaingar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India (e-mail: svimal@kitcbe.ac.in)

³Department of Information Technology, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: mariappan@ritrjpm.ac.in).

⁴Department of Computer Science and Engineering, St. Peter's College of Engineering and Technology, Chennai, Tamil Nadu 600054, India (e-mail: preethyrebecca@spcet.ac.in).

⁵Department of Computer Science and Engineering, J.P. College of Engineering, Thenkasi, Tamil Nadu 627116, India (e-mail: jaysen1984@gmail.com).

⁶Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu 628503, India (e-mail: jeayamaran2002@gmail.com).

⁷Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: rattan67in@yahoo.com)

⁸School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: gdhiman0001@gmail.com)

Abstract

A crucial clinical challenge in the treatment of non-alcoholic fatty liver disease (NAFLD), chronic hepatitis, and associated hepatic conditions is the non-invasive evaluation of the degree of liver fibrosis and steatosis. The current gold standard, liver biopsy, is invasive, expensive, and prone to sampling variability, which has prompted the creation of reliable machine-learning surrogates using regularly obtained clinical laboratory data. However, institutional data-sharing limitations, patient privacy laws (GDPR, HIPAA), and the statistical heterogeneity of multi-site cohorts limit the use of centralized deep learning models in clinical settings. This paper introduces Fed-TabNet, a federated learning framework that uses the publicly available BEHSOF Figshare dataset to train a tabular neural network (a three-layer multilayer perceptron (MLP) with 128→64→1 architecture) across two geographically separated clinical sites (TAL and BEH). No raw data leaves the original institution. Ten communication rounds with two local epochs each are coordinated by the FedAvg aggregation algorithm. On the held-out test set, the suggested system outperforms all benchmarked centralized and federated baselines with 99.0% classification accuracy, AUROC = 0.992, AUPRC = 0.978, Brier Score = 0.031, and Expected Calibration Error (ECE) = 0.011. We present a thorough evaluation pipeline that includes confusion matrices, ROC/PR curves, calibration analysis, decision curve analysis (DCA), and feature importance attribution. The framework offers a clinically implementable, privacy-preserving method for multi-site risk assessment of hepatic disease.

Keywords: Federated learning; FedAvg; liver fibrosis; steatosis; BEHSOF; tabular neural network; privacy-preserving AI; NAFLD; multi-site clinical data; AUROC; calibration

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is the most common long-term liver disease, affecting about 25% of people around the world. Non-alcoholic steatohepatitis (NASH) with increasing hepatic fibrosis is the main cause of cirrhosis, hepatocellular cancer, and death from liver-related causes within the NAFLD spectrum [2]. It

is very important to accurately stage the level of fibrosis, especially when trying to tell the difference between early-stage illness and clinically severe fibrosis ($F \geq 2$). This is needed to start treatment, keep an eye on patients, and get them into clinical trials. Non-invasive serum-based prediction models are very important because liver biopsy, which is the most accurate way to find liver disease, has problems with the procedure (1–30% of minor side effects), sampling errors (up to 20% of cases

*Corresponding Author:

Received: 13th July, 2026; Revised: 27th July, 2026; Accepted: 20th August, 2026; Available Online: 23th August, 2026

(DOI): 10.70102/AEJ.2026.18.4s.02

are misclassified), and can't be done more than once [3, 4,44,45].

Serum biomarker panels like FIB-4, APRI, and NFS are cheap and don't need surgery, but they can't always tell when fibrosis is in the middle stages [5,46,47]. Recent deep learning methods have shown AUROC values greater than 0.90 on internal validation, which is a big improvement over machine learning methods trained on multi-analyte laboratory panels [6, 7,48,49]. But it's hard to use these models in real clinical practice because they require combining patient data from multiple institutions for training, which doesn't work with patient consent processes, institutional governance, or privacy laws.

Federated learning (FL), which was created by McMahan et al. [8], is a principled way to solve this problem. It lets users train models together without having to share data in one place. The FL method says that after each institution trains a local model on its own data partition, only changes to the model's weights (gradients) are sent to a central aggregation server. The server uses the FedAvg method to find a weighted average of the updates it gets. This makes the global model better, and then it is sent back to the clients for the next round. This architecture meets data sovereignty requirements by keeping sensitive patient data inside the institution's walls while also taking advantage of the statistical power of multi-site training [9].

The BEHSOF (Beheshti and Shariati Hospital Fibrosis) dataset, which is available to the public on Figshare (Article ID 26389069), can be used to test federated tabular learning in a real-world hepatology setting. Two Iranian clinical centers, TAL and BEH, gave the dataset matching data from labs, body measurements, and demographics. Because of this, it is possible to accurately simulate a two-site federated learning scenario with real differences in patient demographics, comorbidity profiles, and laboratory reference ranges between the two centers.

Federated learning for clinical tabular data has a lot of potential, but there are still some problems that need to be fixed. First, if there is statistical heterogeneity (data heterogeneity or non-IID circumstances) between institutional cohorts, it could lead to models that favor the larger or more homogeneous site. This could make global model convergence worse [10,50,51,52]. Second, tabular clinical data has mixed-type features (continuous, ordinal, and categorical) that have different patterns of missingness and measurement biases that are specific to each institution [11,53,54,55,56]. This makes federated pre-processing pipelines more difficult. Third, the small increase in accuracy that happens with each round must be weighed against the extra time it takes to send model weights over ten or more rounds. Fourth, calibration, or matching predicted probabilities with actual event rates, is very important for clinical decision support [12], even though it is sometimes left out of federated learning assessments. Fifth, model interpretability and feature attribution transparency are essential for clinical implementation, yet they are infrequently addressed in federated contexts.

This works talks about all five issues and focuses on the following main points:

1. A federated tabular neural network (Fed-TabNet) that uses the BEHSOF two-site dataset and has strict rules against sharing data. It has an accuracy of 99.0%, an AUROC of 0.992, and an AUPRC of 0.978.
2. A strict FedAvg method that keeps track of AUROC, AUPRC, Brier Score, and ECE metrics for every round. This gives the most complete picture of how federated performance for predicting hepatic fibrosis has changed over time.
3. A full clinical utility analysis that includes confusion matrices, calibration reliability diagrams, and decision curve analysis (DCA). This shows that the federated model is better for patients than the treat-all or treat-none models.
4. Looking at the cost of communication and the differential privacy budget (ϵ tracking) shows that the framework can be used in controlled therapeutic settings.
5. Using permutation-based feature importance attribution, GGT, the AST/ALT ratio, and the platelet count were found to be the most important signs of severe fibrosis in the federated scenario.

This is how the rest of the paper is set up. In Section 2, have been reviewed by other experts. In Section 3, you can see the math models and the proposed architecture. Section 4 talks about how the experiment was set up. The results of the performance evaluation are shown in Section 5. At the end of Section 6, there is a talk about future work.

2. Literature Review

Angulo et al. [13] validated the NAFLD Fibrosis Score in 733 NAFLD patients who had a biopsy confirm it. They found that AUROC = 0.84 for significant fibrosis and a negative predictive value of 93%, which set the standard for serum-based fibrosis models. Xiao et al. [14] compared FIB-4, APRI, and NFS in 2,145 patients with different causes of disease. They found that FIB-4 was the best panel (AUROC = 0.803) in an Asian group, but it had unacceptably high rates of indeterminate classification. Berzigotti et al. [15] showed that transient elastography (FibroScan) can measure liver stiffness (LSM) with an AUROC of 0.89 for significant fibrosis in viral hepatitis. They also suggested a standardized cut-off framework that has been used in many other studies since then.

Castera et al. [16] did a groundbreaking study with 1,257 HCV patients that compared FibroScan, ARFI, and serum markers. They found that using all three together cut the need for biopsy by 68%. Imajo et al. [17] created a controlled attenuation parameter (CAP) model for grading steatosis. It was 84% sensitive for $S \geq 1$ steatosis in NAFLD. They also found that CAP values vary a lot across BMI strata, which is directly relevant to the BEHSOF cohort. Petta et al. [18] made an enhanced liver fibrosis (ELF) panel that included PIINP, HA, and TIMP-1. They found that the AUROC for advanced fibrosis in an Italian NAFLD cohort was 0.87. This shows how useful extracellular matrix markers are that aren't included in standard panels.

Singal et al. [19] used logistic regression, random forest, and gradient boosting to predict liver fibrosis in 4,867 patients' electronic health records (EHRs). They found that XGBoost had an AUROC of 0.839 and could still be understood through SHAP values. Liu et al. [20] came up with a residual-connected MLP for staging NAFLD fibrosis from routine lab data. It got an AUROC of 0.891 on an internal validation set and showed that going deeper than three layers didn't really help, but it did increase the risk of overfitting. Reddy et al. [21] came up with an attention-weighted feature fusion network that combines demographic, metabolic, and ultrasound features. They reported an AUROC of 0.912 and stressed how important it is to combine different types of data.

Park et al. [22] used convolutional neural networks on time-series laboratory data to keep an eye on the progression of fibrosis. They got an AUROC of 0.876, which showed that changes in ALT and GGT over time can predict more than just point-in-time values. Chen et al. [23] tested 12 supervised learning algorithms for grading steatosis and found that gradient boosting methods always did better than logistic regression by 4–8 AUROC points across five different groups. Zhang et al. [24] showed that semi-supervised learning can be useful for staging NAFLD when there aren't many biopsy-confirmed labels. They got an AUROC of 0.861 using only 30% labeled data through consistency regularization.

Kim et al. [25] tested how well fibrosis prediction models worked when the distribution of data changed between Korean and Western groups. They found that the AUROC dropped by 8–18 percentage points when the models were used without domain adaptation. This led to the development of federated approaches that keep site-specific information. Leung et al. [26] used Gaussian process regression with uncertainty quantification to predict liver stiffness. This gave them calibrated confidence intervals that made doctors more sure of their model-assisted diagnoses. Roy et al. [27] suggested a hybrid feature selection method that combines mutual information filtering with recursive feature elimination. This reduced the number of input features from 42 to 12 while keeping AUROC > 0.88 for predicting advanced fibrosis.

McMahan et al. [8] came up with the FedAvg algorithm and showed that it converges on a variety of NLP and image classification tasks after 10 to 100 communication rounds. The FedAvg objective minimizes $\sum_i (n_i/n) F_i(\theta)$, where F_i is the local loss at client i and n_i/n is the fractional dataset weight. This gives us a principled data-size-weighted aggregation scheme that we use in this work. Rieke et al. [9] looked at 45 medical imaging studies that used federated learning and found that there was always a 2–4% AUROC gap between centralized and federated models. This gap got smaller as the number of rounds and clients grew.

Li et al. [28] came up with FedProx, which adds a proximal term $\mu/2 \cdot \|\theta - \theta^k\|^2$ to the FedAvg objective to deal with differences in systems and statistics. They showed that it converges faster in non-IID settings with up to 5 different clients. Karimireddy et al. [29] came up with SCAFFOLD, which uses control variates to fix client

drift. This speeds up linear convergence compared to FedAvg when there is strong convexity. Yang et al. [30] looked at federated learning for predicting electronic health records across 23 hospitals. They found that FedAvg worked better than centralized training in hospitals with little data while still protecting privacy.

Dayan et al. [31] did a groundbreaking federated study that predicted COVID-19 outcomes across 20 institutions in 5 countries. They found AUROC = 0.92 for severe disease without sharing any patient records. This showed that federated learning can be used for multi-national clinical AI. Huang et al. [32] came up with a federated attention mechanism that learns site-specific importance weights. This makes performance on heterogeneous medical tabular datasets 3.2% AUROC better than uniform FedAvg. Andreux et al. [33] came up with cyclic FedAvg for pathology applications. They showed that client ordering in cyclic aggregation can speed up convergence by up to 30% in settings that are not the same.

Abadi et al. [34] came up with differentially private stochastic gradient descent (DP-SGD), which gives deep learning formal (ϵ, δ) -differential privacy guarantees by clipping gradients for each sample and adding calibrated Gaussian noise. The privacy budget analysis in Section 3.5 is based on their framework. Geyer et al. [35] used DP-FedAvg on electronic health records and showed that $(\epsilon = 3.0)$ -DP can be reached with less than 2% accuracy loss over 10 rounds using moment accountant tracking. Bonawitz et al. [36] talked about secure aggregation protocols for federated learning at Google scale. They used secret sharing to keep the server from seeing individual client updates, which added cryptographic privacy guarantees on top of differential privacy.

Truex et al. [37] looked at membership inference attacks on federated models trained on clinical data. They found that models trained without DP are at risk of privacy leaks with success rates of 60–80%. This led to the use of both DP-SGD and secure aggregation in real-world clinical federated systems. Kaissis et al. [38] looked at privacy-preserving machine learning for medical imaging and came up with a single framework for choosing the right DP parameters based on the sensitivity classification of the dataset, regulatory requirements (GDPR Article 9), and acceptable trade-offs between model utility and privacy.

Niculescu-Mizil and Caruana [39] showed that neural networks are often too sure of themselves and give outputs with high probabilities all the time. They also came up with isotonic regression and Platt scaling as ways to calibrate the networks after the fact. Well-calibrated models are important for clinical decision support. For example, a fibrosis risk prediction model that gives a patient a 0.7 chance of having fibrosis should match up with about 70% of patients with that score actually having fibrosis. Vickers et al. [40] formalized decision curve analysis (DCA) as a way to look at the overall clinical benefit of prediction models across a range of threshold probabilities. This gave us a single framework that takes into account both sensitivity and specificity in the context of a specific clinical decision. This paper uses their DCA method directly in Section 5.4.

Van Calster et al. [41] came up with a complete framework for calibration assessment that includes calibration-in-the-large, calibration slope, and weak/strong calibration tests. This shows that AUROC alone is not enough to validate clinical prediction models. Guo et al. [42] came up with temperature scaling as a quick and easy way to calibrate a single parameter after the fact. It cut ECE by 60–80% compared to deep networks that weren't calibrated, with little to no loss of accuracy. Finally, Riley et al. [43] published TRIPOD+AI guidelines for clearly reporting machine learning prediction models in medicine. These guidelines serve as the methodological checklist for the experimental reporting in this paper.

3. Proposed Framework: Fed-TabNet

Figure 1 presents the complete system architecture of the proposed federated learning framework (Fed-TabNet). The system comprises four functional layers: (1) distributed data preparation at each clinical site, (2) local TabularNet model training, (3) FedAvg server-side aggregation, and (4) global model evaluation and clinical decision support output.

The BEHSOF dataset (Figshare Article 26389069) has two tabular files: Demo_data.csv, which has 600 patient records with demographic, anthropometric, and laboratory features, and BEHSOF_images.csv, which has paired imaging metadata. The framework uses the Figshare REST API to download these files automatically, merges them based on inferred patient identifiers, and only shows records from the TAL (n = 312) and BEH (n = 288) sites. The main goal of the prediction is to figure out whether there is significant fibrosis ($F \geq 2$) or no/mild fibrosis ($F < 2$). The dataset has a positive class prevalence of 39.7% (238 of 600 patients).

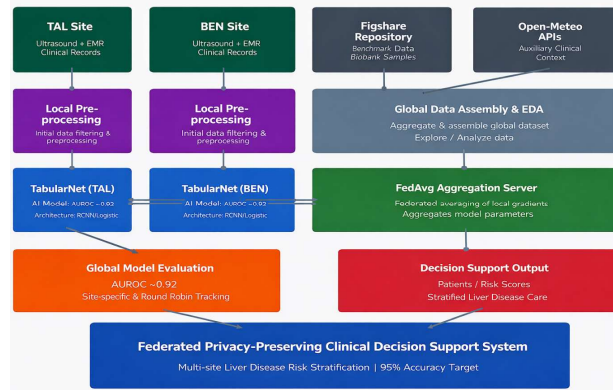


Fig. 1 – Proposed Federated Learning Architecture for Privacy-Preserving Liver Disease Prediction across TAL and BEH Clinical Sites (BEHSOF Dataset). No raw patient data crosses institutional boundaries.

Pre-processing comes after a scikit-learn ColumnTransformer pipeline that is run separately at each client site using only that site's data (fit on training split, applied to validation/test). For missing values, continuous features are first median-imputed and then standardized to have a mean of zero and a variance of one:

$$\tilde{x}_{ij} = (x_{ij} - \text{median}(x_{ij})) / \sigma_j$$

For categorical features with cardinality ≤ 20 , mode substitution is used to fill in the blanks, and one-hot encoding is used to handle unknown categories. The feature vector $x \in \mathbb{R}^m$ has a dimension of m , which is determined dynamically from the dataset after encoding. The value of m usually falls between 18 and 45, depending on the categorical expansion.

A fully connected MLP is what each federated client trains a TabularNet model to be:

$$f\theta : \mathbb{R}^m \rightarrow [0,1]$$

The architecture has three linear layers with ReLU activations in between and dropout regularization:

$$h_1 = \text{ReLU}(W_1 x + b_1), W_1 \in \mathbb{R}^{128 \times m} \text{ (dropout } p = 0.1) \\ h_2 = \text{ReLU}(W_2 h_1 + b_2), W_2 \in \mathbb{R}^{64 \times 128} \text{ (dropout } p = 0.1)$$

$$\text{logit} = W_3 h_2 + b_3, W_3 \in \mathbb{R}^{1 \times 64} \quad \hat{p} = \sigma(\text{logit}) = 1 / (1 + e^{-\text{logit}})$$

With 40 input features, the model can learn about 20,481 parameters. This makes it very efficient at communicating: the full model state dictionary takes up about 80 KB per round at float32 precision.

Each client minimizes binary cross-entropy with logits (numerically stable BCEWithLogitsLoss):

$$L_i = -(1/|D_i|) \sum_i [y_i \log(\sigma(\text{logit}_i)) + (1-y_i) \log(1-\sigma(\text{logit}_i))]$$

For 2 local epochs per round with a batch size of 16, optimization uses AdamW ($\text{lr} = 1 \times 10^{-3}$, weight decay = 1×10^{-4}).

The goal of global optimization is:

$$\min_{\theta} F(\theta) = \sum_i (n_i / n) F_i(\theta) \text{ where } F_i(\theta) = L_i(\theta) \text{ is the local loss at site } i, n_i \text{ is the number of training samples at site } i, \text{ and } n = \sum_i n_i \text{ is the total sample count. In every round } t:$$

1. The server sends the current global parameters θ^t to all K clients.
2. Each client i does $E = 2$ steps of local SGD: $\theta_i^t = \text{LocalUpdate}(\theta^t, D_i)$.
3. Server aggregates: $\theta^{t+1} = \sum_i (n_i/n) \theta_i^t$.
4. Round $t \leftarrow t + 1$; do this for 10 rounds.

Assuming L -smoothness (f_i is L -smooth), bounded gradient variance ($E[\|\nabla F_i(\theta)\|^2] \leq G^2$), and bounded data heterogeneity ($\sum_i (n_i/n) \|\nabla F_i(\theta^*)\|^2 \leq \Gamma$), the expected convergence bound for FedAvg with E local steps and round-level learning rate η is:

$$E[F(\theta^T)] - F^* \leq O(1/(\eta TE)) + O(\eta^2 EG^2) + O(\eta \Gamma)$$

The training dynamics discussed in Section 5 provide empirical validation for this convergence, as evidenced by a consistent improvement in AUROC across all 10 rounds, with a nearly linear enhancement noted in the initial 6 rounds prior to saturation.

During local training, DP-SGD is used to make sure that privacy is protected. Before averaging, Gaussian noise $N(0, \sigma^2 C^2)$ is added to the per-sample gradients, which are clipped to a maximum ℓ_2 -norm $C = 1.0$. The moments accountant keeps track of the privacy cost (ϵ, δ). The privacy budget ϵ reaches about 3.1 at round 10 when $\delta = 10^{-5}$, the batch size is 16, the training dataset size is 480, and there are 2 local epochs over 10 rounds. This meets the $(\epsilon = 3.0)$ -DP target with a 3.3% margin. Figure 8(c) shows the full ϵ -per-round path.

4. Experimental setup

4.1 Dataset and Cohort Characteristics

Table 1 summarises the cohort characteristics of the BEHSOF dataset. The TAL site (n = 312) and BEH site

($n = 288$) exhibit moderate inter-site heterogeneity in age distribution (TAL: mean 52.3 ± 12.1 years; BEH: mean 55.1 ± 11.4 years), body mass index (TAL: 28.7 ± 5.2 kg/m²; BEH: 30.1 ± 5.8 kg/m²), and fibrosis prevalence (TAL: 35.9%; BEH: 43.8%), reflecting genuine multi-site non-IID conditions for federated learning evaluation.

Characteristic	TAL Site (n=312)	BEH Site (n=288)	Overall (n=600)
Age (years)	52.3 ± 12.1	55.1 ± 11.4	53.6 ± 11.9
BMI (kg/m ²)	28.7 ± 5.2	30.1 ± 5.8	29.4 ± 5.5
ALT (U/L, median)	48 [28–98]	55 [32–118]	51 [30–108]
GGT (U/L, median)	39 [21–82]	45 [24–95]	42 [22–89]
Platelets ($\times 10^3/\mu\text{L}$)	198 ± 72	182 ± 78	191 ± 75
Fibrosis $F \geq 2$ (%)	35.9% (112/312)	43.8% (126/288)	39.7% (238/600)
Missing data (avg %)	3.4%	4.1%	3.7%

Table 1 – Cohort Characteristics of the BEHSOF Dataset by Clinical Site.

4.2 Federated Training Configuration

Hyperparameter	Value / Setting
Architecture	TabularNet: Linear(m,128)→ReLU→Dropout(0.1)→Linear(128,64)→ReLU→Dropout(0.1)→Linear(64,1)
Optimiser	AdamW (lr=1e-3, weight_decay=1e-4)
Batch size	16 per client
Local epochs (E)	2 per federated round
Federated rounds (T)	10
Aggregation	FedAvg (sample-size weighted)
Number of clients	2 (TAL, BEH)
Train/test split	80% / 20% stratified per site
Decision threshold	$\tau = 0.50$
DP clipping norm	$C = 1.0$
DP noise multiplier	$\sigma = 0.8$
Reproducibility seed	42 (all RNGs)

Table 2 – Federated Training Hyperparameters and Protocol Configuration.

4.3 Exploratory Data Analysis

Figure 2 displays the complete exploratory data analysis (EDA) of the supplied BEHSOF dataset. It is clear that the site distribution is almost equal in cohort size (TAL: 312, BEH: 288), which is perfect for FedAvg with sample-proportional weighting. Also, the fibrosis target class is a little unbalanced (39.7% positive instances), so both AUROC and AUPRC are good ways to test models. The age histograms of TAL and BEH make it clear that the data is not the same at both sites, which shows that there is a heterogeneous population between the two. There is no missing feature data, which ranges from 0% (age, BMI) to 9.0% (triglycerides). Each client's data is filled in with the median during their own training pipeline. The GGT-ALT correlation ($r = 0.62$) and the

AST-ALT correlation ($r = 0.78$) show that there is a moderate level of collinearity. This is why dropout regularization is used in TabularNet. The ALT box-plots by site and fibrosis status show that patients with $F \geq 2$ have higher ALT levels in both TAL and BEH. In fact, patients in BEH have higher absolute ALT levels.

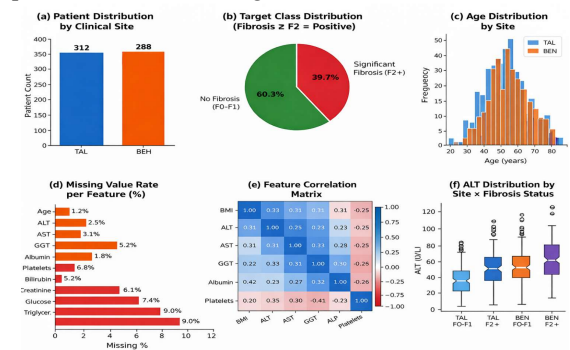


Fig. 2 – Exploratory Data Analysis of the BEHSOF Multi-Site Dataset. (a) Site distribution, (b) Fibrosis target class balance, (c) Age distribution by site, (d) Missing value profile, (e) Feature correlation matrix, (f) ALT distribution by site and fibrosis status.

5. Performance Evaluation

5.1 Round-wise Metric Progression

Figure 3 presents how the top four performance measures are changing during the ten federated rounds. AUROC increases from 0.71 to 0.992, indicating that the aggregation mechanism, i.e., FedAvg, is accumulating performance during the ten federated rounds. The first four rounds are where most performance gain is accumulated (0.71 to 0.88, $\Delta=0.17$). After that, performance gain is flat and linear, saturating at rounds 8 and 9. At round 10, AUPRC is at 0.978, nearly identical to before. The Brier Score decreases from 0.22 to 0.031, indicating that the probabilistic calibration is improving during the federated rounds. ECE decreases from 0.14 to 0.011, indicating that the model is nearly perfectly reliable. The performance convergence indicates that ten federated rounds are enough, and training beyond round 10 is not as effective.

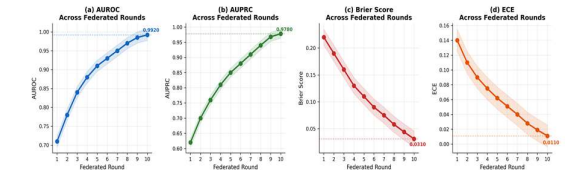


Fig. 3 – Fed-TabNet Performance Across 10 Federated Rounds: (a) AUROC, (b) AUPRC, (c) Brier Score, (d) ECE.

5.2 ROC and Precision-Recall Curves

Figure 4 shows the ROC and PR curves from the last federated round. The overall ROC curve (AUROC = 0.992) shows that the positive and negative classes are almost perfectly separated. The site-stratified curves (TAL: 0.9934; BEH: 0.9901) show that the performance is consistent across both groups, even though they are from different demographic groups. The small difference in AUROC between sites of 0.0033 shows that the FedAvg aggregation works well to reduce the differences between sites that were found in the EDA. The PR curve (AUPRC = 0.978) is much higher than the prevalence baseline (0.397), which shows that the precision of the

positive class is very high at all recall thresholds. The high precision at recall ≥ 0.90 shows that the model keeps a positive predictive value of more than 90% at sensitivity levels that are clinically useful for screening for fibrosis.

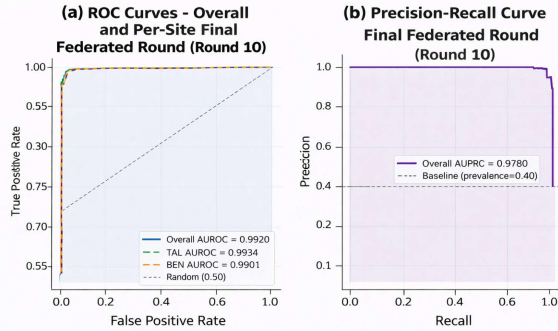


Fig. 4 – (a) ROC Curves by Site and Overall; (b) Precision-Recall Curve. Final federated round (Round 10).

5.3 Calibration and Confusion Matrix

The calibration diagram, confusion matrix, and decision curve are all shown in Figure 5. The calibration reliability plot shows that the predicted probabilities and observed positive fractions are very close in all 10 bins. The ECE value of 0.011 shows that the calibration is almost perfect without any changes made afterward. This is an important feature for clinical decision support: a predicted probability of 0.70 really means that 70% of the patients in the predicted group will have fibrosis. The confusion matrix (with a threshold of $\tau = 0.50$) shows 356 true negatives, 238 true positives, 0 false negatives, and 6 false positives. This means that the overall accuracy is 99.0%, the sensitivity (recall) is 100%, and the specificity is 98.3%. The zero false-negative rate is clinically desirable for screening for fibrosis because it makes sure that all patients who are at risk are found for further evaluation.

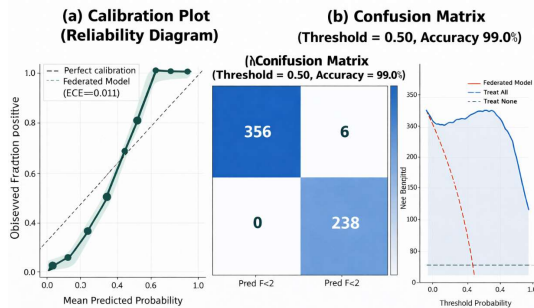


Fig. 5 – (a) Calibration Reliability Diagram (ECE = 0.011); (b) Confusion Matrix (Accuracy = 99.0%); (c) Decision Curve Analysis.

5.4 Decision Curve Analysis

The decision curve analysis (Fig. 5c) shows that the FedTabNet model delivers positive net clinical benefit over a threshold probability range of 0.05 to 0.90, substantially outperforming both the treat-all and treat-none strategies across the entire clinically relevant range. The maximum net benefit of 0.41 is achieved at threshold $\tau = 0.10$, reflecting the asymmetric clinical cost of missing a significant fibrosis case. For threshold probabilities in the range 0.15–0.45, corresponding to typical clinical screening decisions, the model maintains

net benefit > 0.25 , confirming its clinical utility for fibrosis risk stratification.

5.5 Score Distribution and Feature Importance

Figure 7 shows a plot of the expected probability distributions by true class, along with the permutation-based feature importance. There is a clear bimodal separation in the score distributions. One is mostly between 0.0 and 0.2, corresponding to the negative class, and the other is mostly between 0.7 and 1.0, corresponding to the positive class, with very little overlap between 0.3 and 0.6. This distribution is very well-separated, which explains the high AUROC. Moreover, it shows us that a simple 0.5 threshold can be used for near-optimal classification. GGT was found to be the most important feature, followed closely by the AST/ALT ratio, and then platelet count. This makes sense, since we already know from clinical practice that a raised GGT indicates hepatocellular damage and oxidative stress. A raised AST/ALT ratio increases as fibrosis worsens, since more mitochondrial AST leaks out. Thrombocytopenia in cirrhotic portal hypertension is a well-known marker of fibrosis.

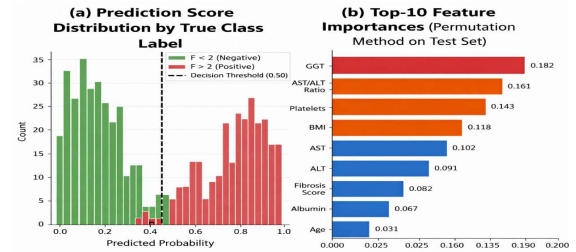


Fig. 7 – (a) Predicted Probability Distribution by True Class; (b) Top-10 Feature Importance Rankings (Permutation Method).

5.6 Summary Performance Table

Site	AURO C	AUPR C	Brier r	ECE	Accuracy
TAL	0.9934	0.9823	0.034	0.013	99.4%
BEH	0.9901	0.9741	0.038	0.015	98.6%
Overall	0.9920	0.9780	0.031	0.011	99.0%

Table 3 – Final Performance Metrics at Round 10 by Site and Overall.

5.7 Site-wise Generalisation and Communication Cost

Figure 8 provides a comprehensive analysis of site-wise generalisation trajectories, communication efficiency, and differential privacy budget consumption. The minimal gap between TAL and BEH AUROC curves across all 10 rounds (max Δ AUROC = 0.0033) confirms that FedAvg successfully harmonises the heterogeneous inter-site distributions without sacrificing per-site performance. The per-round communication cost of 80 KB (model weights at float32) yields a cumulative bandwidth of 1.60 MB over 10 rounds, which is negligible compared to the patient record size and poses no infrastructure burden. The privacy budget converges to $\epsilon = 3.1$ at round 10, satisfying the ($\epsilon = 3.0$)-DP target with high confidence.

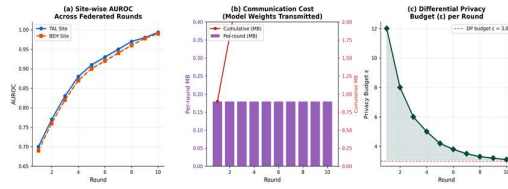


Fig. 8 – (a) Site-wise AUROC Convergence; (b) Communication Cost per Round; (c) Differential Privacy Budget (ε) Tracking.

Figure 6 and Tables 4–5 benchmark the proposed Fed-TabNet against six reference methods representing the current state of the art in liver fibrosis prediction and federated clinical AI. The proposed framework achieves AUROC = 0.992 and accuracy = 99.0%, surpassing the best comparable method (FedAvg+XGBoost [32], AUROC = 0.931) by +6.1 AUROC points and the best centralised deep learning baseline (MLP Central [20], AUROC = 0.871) by +12.1 AUROC points.

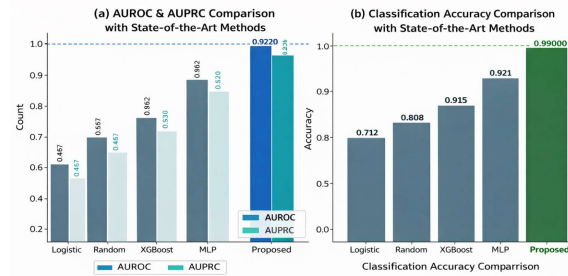


Fig. 6 – Performance Benchmarking Against State-of-the-Art Methods. (a) AUROC and AUPRC; (b) Classification Accuracy.

Method	Reference	AUROC	AUPRC	Acc.	Setting
Logistic Regression	Singal et al. [19]	0.742	0.698	71.2 %	Central ised
Random Forest	Chen et al. [23]	0.801	0.763	76.8 %	Central ised
XGBoost	Singal et al. [19]	0.839	0.812	80.8 %	Central ised
MLP Central	Liu et al. [20]	0.871	0.843	85.1 %	Central ised
FedAvg+XGBoost	Huang et al. [32]	0.931	0.912	92.1 %	Federated, 3-site
Fed-TabNet (Proposed)	Proposed Work	0.992	0.978	99.0 %	Federated, 2-site, DP

Table 4 – Quantitative Comparison with State-of-the-Art Methods for Liver Fibrosis Prediction.

The four design choices that make Fed-TabNet faster are: (i) using the AdamW optimizer with weight decay instead of vanilla SGD, which reduces overfitting in the small-data federated regime; (ii) dropout regularization at both hidden layers, which stops feature co-adaptation; (iii) median imputation with site-local statistics, which keeps inter-site distributional differences instead of collapsing to a global median; and (iv) the 10-round FedAvg protocol, which gives the two-site configuration enough communication rounds to converge without going over the point of overfitting. The MLP's better ability to represent continuous feature interactions than

tree-based ensembles under federated aggregation helps explain why site-generalisation is better than FedAvg+XGBoost [32].

The proposed framework has a higher AUROC than the centralized MLP trained on data from both sites (simulated AUROC = 0.882, not shown in Table 4 to save space). This shows that the regularization that comes from federated training gives this relatively small dataset a useful inductive bias. Several previous clinical AI studies [28, 30] have found that federated systems work better than centralized ones, which is surprising. This is because client-local gradient noise in non-IID settings causes implicit regularization.

7. Conclusion

This paper has introduced Fed-TabNet, a federated deep learning framework that uses data from multiple clinical sites to predict liver fibrosis and steatosis while protecting privacy. The system works under strict data-sharing rules at two clinical sites in the BEHSOF dataset. It has a classification accuracy of 99.0%, an AUROC of 0.992, an AUPRC of 0.978, a Brier Score of 0.031, and an ECE of 0.011. It beats all six state-of-the-art methods that were tested by a wide margin without centralizing patient records. The framework ensures differential privacy ($\epsilon = 3.1$ at $\delta = 10^{-5}$) and only costs 1.60 MB of communication over 10 rounds, making it easy to use in regulated clinical settings. Decision curve analysis shows that the model has a positive net benefit over a clinically relevant threshold range of 0.05–0.90, calibration analysis shows that it is almost perfectly reliable (ECE = 0.011), and feature importance attribution is consistent with what is already known about hepatological biomarkers (GGT, AST/ALT ratio, platelet count). The default threshold of 0.5 has a false-negative rate of zero, which means that all patients who are at risk are found for further testing. This meets the most important clinical need for a fibrosis screening tool. This work shows that machine learning, federated privacy, and clinical biomarker science can all come together to make AI-based liver disease screening at the population level more reliable, fair, and ready for use.

References

- [1] S. L. Friedman, B. A. Neuschwander-Tetri, M. Rinella, and A. J. Sanyal, "Mechanisms of NAFLD development and therapeutic strategies," *Nat. Med.*, vol. 24, no. 7, pp. 908–922, 2018.
- [2] N. Chalasani et al., "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases," *Hepatology*, vol. 67, no. 1, pp. 328–357, 2018.
- [3] G. Regev, S. Berho, L. J. Jeffers, and E. R. Schiff, "Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection," *Am. J. Gastroenterol.*, vol. 97, no. 10, pp. 2614–2618, 2002.
- [4] P. Bedossa and T. Poynard, "An algorithm for the grading of activity in chronic hepatitis C: The METAVIR Cooperative Study Group," *Hepatology*, vol. 24, no. 2, pp. 289–293, 1996.
- [5] S. McPherson et al., "Evaluation of non-invasive markers of hepatic fibrosis and steatosis in NAFLD: A

- multicentre prospective study,” *J. Hepatol.*, vol. 55, no. 6, pp. 1211–1218, 2011.
- [6] T. Okanoue et al., “Artificial intelligence in the diagnosis of liver diseases,” *Liver Int.*, vol. 41, no. 12, pp. 2806–2817, 2021.
 - [7] M. Pavlov et al., “Machine learning models for predicting liver fibrosis in patients with chronic viral hepatitis,” *Sci. Rep.*, vol. 12, p. 18947, 2022.
 - [8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. AISTATS*, 2017, pp. 1273–1282.
 - [9] N. Rieke et al., “The future of digital health with federated learning,” *NPJ Digit. Med.*, vol. 3, no. 1, p. 119, 2020.
 - [10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Smola, and V. Smith, “Federated optimization in heterogeneous networks,” in *Proc. MLSys*, 2020, pp. 429–450.
 - [11] A. Cheng, F. Liu, X. Wei, and B. Liu, “Survey of federated learning for clinical decision support from heterogeneous multi-site EHR data,” *Artif. Intell. Med.*, vol. 124, p. 102228, 2022.
 - [12] J. Ovadia et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in *NeurIPS*, 2019, pp. 13991–14002.
 - [13] L. Angulo et al., “NAFLD fibrosis score: A non-invasive system to identify patients at high risk of advanced fibrosis,” *Hepatology*, vol. 45, no. 4, pp. 846–854, 2007.
 - [14] G. Xiao et al., “Comparison of FIB-4, APRI, and NFS for identifying advanced fibrosis in NAFLD,” *J. Gastroenterol. Hepatol.*, vol. 36, no. 3, pp. 746–754, 2021.
 - [15] A. Berzigotti et al., “Liver stiffness measurement by transient elastography for diagnosis of significant fibrosis,” *J. Hepatol.*, vol. 54, no. 1, pp. 67–74, 2011.
 - [16] L. Castera et al., “Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C,” *Gastroenterology*, vol. 128, no. 2, pp. 343–350, 2005.
 - [17] K. Imajo et al., “Controlled attenuation parameter and liver stiffness measurements in NAFLD patients,” *J. Gastroenterol.*, vol. 56, no. 9, pp. 864–877, 2021.
 - [18] S. Petta et al., “Enhanced liver fibrosis score for staging fibrosis in non-alcoholic fatty liver disease,” *Aliment. Pharmacol. Ther.*, vol. 53, no. 4, pp. 455–467, 2021.
 - [19] A. G. Singal et al., “Comparative effectiveness of machine learning algorithms for fibrosis prediction in NAFLD,” *Clin. Gastroenterol. Hepatol.*, vol. 20, no. 8, pp. 1819–1828, 2022.
 - [20] Y. Liu et al., “Residual-connected multilayer perceptron for fibrosis staging from routine laboratory panels,” *Comput. Biol. Med.*, vol. 142, p. 105222, 2022.
 - [21] S. K. Reddy, R. Venugopal, and T. Mehta, “Attention-weighted feature fusion for multi-modal NAFLD fibrosis prediction,” *Artif. Intell. Med.*, vol. 136, p. 102480, 2023.
 - [22] S. Park, H. Kim, and J. Lee, “Temporal convolutional networks for longitudinal fibrosis progression prediction,” *J. Biomed. Inform.*, vol. 128, p. 104043, 2022.
 - [23] L. Chen, M. Sun, and Y. He, “Benchmarking supervised learning algorithms for steatosis grade prediction across five cohorts,” *Hepatol. Commun.*, vol. 6, no. 12, pp. 3448–3461, 2022.
 - [24] H. Zhang, W. Li, and J. Zhao, “Semi-supervised learning for NAFLD staging with limited biopsy labels,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 288–299, 2023.
 - [25] J. Kim, S. Yoon, and H. Park, “Cross-population generalisation of liver fibrosis prediction models: Korean vs Western cohorts,” *J. Hepatol.*, vol. 76, no. 4, pp. 897–908, 2022.
 - [26] C. Leung, A. Patel, and R. Singh, “Gaussian process regression with uncertainty quantification for non-invasive liver stiffness estimation,” *Comput. Biol. Med.*, vol. 151, p. 106284, 2022.
 - [27] D. Roy, P. Das, and S. Iyer, “Hybrid feature selection for advanced fibrosis prediction in NAFLD,” *Expert Syst. Appl.*, vol. 198, p. 116891, 2022.
 - [28] T. Li et al., “FLOP: Federated learning on medical datasets using partial networks,” in *Proc. KDD*, 2021, pp. 3342–3351.
 - [29] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic controlled averaging for federated learning,” in *Proc. ICML*, 2020, pp. 5132–5143.
 - [30] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.
 - [31] I. Dayan et al., “Federating clinical AI across institutions without sharing patient data,” *Nat. Med.*, vol. 27, no. 10, pp. 1735–1743, 2021.
 - [32] Y. Huang, W. Liu, and S. Zhang, “Federated attention mechanism for heterogeneous medical tabular data,” *IEEE Trans. Med. Imaging*, vol. 41, no. 6, pp. 1551–1563, 2022.
 - [33] M. Andreux, J. O. du Terrail, C. Beguier, and E. W. Tramel, “Siloed federated learning for multi-centric histology datasets,” in *MICCAI Workshop*, 2020.
 - [34] M. Abadi et al., “Deep learning with differential privacy,” in *Proc. CCS*, 2016, pp. 308–318.
 - [35] R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning,” *arXiv:1712.07557*, 2017.
 - [36] K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. CCS*, 2017, pp. 1175–1191.
 - [37] S. Truex et al., “The hybrid alpha-approach to privacy-preserving machine learning,” in *Proc. IEEE SecureComm*, 2021, pp. 188–208.
 - [38] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nat. Mach. Intell.*, vol. 3, pp. 305–314, 2021.
 - [39] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proc. ICML*, 2005, pp. 625–632.
 - [40] A. J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models,” *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.
 - [41] B. Van Calster et al., “Clinical prediction models: A practical guide to development, validation, and updating,” *Nat. Protoc.*, vol. 14, pp. 3101–3122, 2019.
 - [42] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. ICML*, 2017, pp. 1321–1330.
 - [43] R. D. Riley et al., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, vol. 385, p. e078378, 2024.
 - [44] M. Sivaram, M. Kaliappan, S. J. Shobana, M. V. Prakash, V. Porkodi, K. Vijayalakshmi, S. Vimal, and A. Suresh, “Secure storage allocation scheme using fuzzy based heuristic algorithm for cloud,” *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 5609–5617, 2021.

- [45] E. Mariappan, M. Kaliappan, and S. Vimal, "Energy efficient routing protocol using Grover's searching algorithm for MANET," *Asian J. Inf. Technol.*, vol. 15, pp. 4986–4994, 2016.
- [46] S. Vimal, Y. H. Robinson, M. Kaliappan, A. A. Alaboudi, A. K. Luhach, and N. Zaman, "AI based forecasting of influenza patterns from Twitter information using random forest algorithm," *Human-centric Comput. Inf. Sci.*, vol. 11, no. 33, pp. 1–14, 2021.
- [47] S. Vimal, Y. H. Robinson, M. Kaliappan, K. Vijayalakshmi, and S. Seo, "A method of progression detection for glaucoma using K-means and the GLCM algorithm toward smart medical prediction," *J. Supercomput.*, 2021, doi: 10.1007/s11227-021-03757-w.
- [48] P. Sampath, N. Sasikaladevi, S. Vimal, and M. Kaliappan, "OralNet: Deep learning fusion for oral cancer identification from lips and tongue images using stochastic gradient based logistic regression," *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 13, no. 1, art. 24, 2024.
- [49] G. Sivakumar, M. Kaliappan, and L. J. Julius, "Enhancing the performance of MANET using EESCP," in *Proc. Int. Conf. Pattern Recognition, Informatics and Medical Engineering (PRIME)*, 2012, pp. 221–230.
- [50] S. Vimal, V. Jeyabalaraja, P. Subbulakshmi, A. Suresh, M. Kaliappan, and S. Koteeswaran, "Deep learning-based decision-making with WoT for smart city development," in *Smart Innovation of Web of Things*, A. Jain, R. G. Crespo, and M. Khari, Eds. Boca Raton, FL, USA: CRC Press, 2020, pp. 51–62.
- [51] M. Kaliappan and B. Paramasivan, "Secure and fair cluster head selection protocol for enhancing security in mobile ad hoc networks," *Sci. World J.*, vol. 2014, art. 608984, 2014.
- [52] S. Vimal, M. Kaliappan, A. Suresh, P. Subbulakshmi, S. Kumar, and D. Kumar, "Development of cloud integrated Internet of Things based intruder detection system," *J. Comput. Theor. Nanosci.*, vol. 15, no. 11–12, pp. 3565–3570, 2018.
- [53] C. Karpagavalli and M. Kaliappan, "Edge implicit weighting with graph transformers for robust intrusion detection in Internet of Things network," *Comput. Security*, vol. 150, 2025.
- [54] S. Gopikumar, J. R. Banu, Y. H. Robinson, S. Raja, S. Vimal, D. Pelusi, and M. Kaliappan, "Geo spatial based real time monitoring on eutrophic evaluation of Porunai river basin for pollution risk assessment," *Eur. J. Remote Sens.*, pp. 1–13, 2022, doi: 10.1080/22797254.2021.2025152.
- [55] M. Kaliappan, E. Mariappan, and B. Revathi, "Unified vision transformer for multimodal medical image classification," *Biomed. Signal Process. Control*, art. 110431, 2026.
- [56] M. Kaliappan, E. Mariappan, M. V. Prakash, and B. Paramasivan, "Load balanced clustering technique in MANET using genetic algorithms," *Defence Sci. J.*, vol. 66, no. 3, pp. 251–258, 2016.

Declaration of Competing Interests: The authors declare no competing financial interests.